Can Gradient Information Be Used to Improve Variational Objective Analysis?

PHILLIP L. SPENCER*

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

JIDONG GAO

Center for Analysis and Prediction of Storms, University of Oklahoma, Norman, Oklahoma

(Manuscript received 29 December 2003, in final form 21 June 2004)

ABSTRACT

A variational scheme for the analysis of scalar variables is developed and compared to two-pass and three-pass versions of the Barnes analysis scheme. The variational scheme, appropriate for diagnostic studies, is similar to a previously developed variational method in that scalar gradient ''observations''—derived directly from the scalar observations—are used in addition to the scalar observations themselves. The current scheme is different in that the cost function does not require analyses of the scalar field and its gradient; it simply requires scalar and gradient observations at their native locations. For the evaluation, randomly selected model gridpoint data are chosen to serve as pseudo-observations for the analysis schemes. By choosing appropriate model gridpoint data to serve as pseudo-observations, artificial data networks can be generated so as to mimic the spatial characteristics of real observational networks.

Results indicate that the proposed variational scheme is superior to both two-pass and three-pass Barnes schemes, increasingly so as the observations become more irregularly spaced. This is true even when the gradient information is not allowed to affect the variational analyses. When the observations are relatively sparse and irregularly distributed, further improvements in the variational analyses occur when the gradient information is properly included within the analysis scheme.

1. Introduction

Procedures for mapping observations that are irregularly distributed in space and/or time to a regular grid—a process called *objective analysis*—have existed and evolved for many decades. Historically, the general purposes for employing objective analysis methods are twofold: to aid in the diagnostic studies of various weather phenomena and to provide the initial conditions for numerical weather forecasts. Successful diagnostic studies usually depend on accurate estimations of spatial derivatives of various quantities (e.g., wind, temperature, humidity, pressure). These spatial derivatives are computed easily from finite-differencing methods which, in turn, require gridded analyses. The use of traditional objective analysis schemes (e.g., Cressman 1959; Barnes 1964, 1973) for providing initial conditions to numerical models has all but disappeared in favor of using more sophisticated schemes that account for the statistical properties of the background field and observations, as well as the relationships between variables that are described by various dynamic and/or continuity equations. Optimal interpolation (OI) and variational methods are examples of such modern, sophisticated analysis schemes. Although of little or no use for present-day numerical forecasting, the use of simple analysis schemes for diagnostic studies continues to this day. Lorenc (1986), Caracena (1987), Parsons and Dudhia (1997), and Spencer et al. (2003, hereafter SSF03) argue that for this purpose, sometimes it is preferable to use a simple analysis scheme rather than modern, sophisticated analysis techniques. The reader is referred to these papers for further discussion and examples.

One of the early ''simple'' analysis schemes involved surface fitting, whereby observations were fit to second- or third-order polynomials (Panofsky 1949; Gilchrist and Cressman 1954). These surface-fitting methods provided reasonable analyses in data-dense regions, but performed less admirably elsewhere. Another simple scheme, the method of successive corrections, was proposed by Bergthórsson and Döös (1955) and made popular by Cressman (1959) and Barnes (1964, 1973), whose schemes remain popular today. The basic idea behind this method is that a background field (also known as a first-guess field) can be successively ad-

* Additional affiliation: NOAA/National Severe Storms Laboratory, Norman, Oklahoma.

Corresponding author address: Phillip L. Spencer, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: phillip.spencer@noaa.gov

justed based on the observations until the analysis appropriately converges to the observations. The background field may be provided by climatology, persistence, or a short-term model forecast.[1] The weight functions that determine an observation's influence upon the value of a particular grid point are empirical and distance dependent. Thus, the Cressman and Barnes schemes represent distance-dependent weighted-averaging objective analysis schemes.

When analysis methods such as these are used to create gridded fields, finite-differencing schemes typically are used to estimate the spatial derivatives. Gridded fields, however, are not a prerequisite for estimating derivatives. In fact, first-order spatial derivatives can be estimated *directly* from the observations themselves, rendering unnecessary the creation of gridded fields for finite-differencing purposes (Bellamy 1949; Endlich and Clark 1963; Ceselski and Sapp 1975; Schaefer and Doswell 1979; Zamora et al. 1987; Doswell and Caracena 1988). Therefore, when gridded fields of derivative quantities are desired, analysts have two options: 1) apply a finite-differencing scheme to the gridded observations, or 2) apply an analysis scheme to the irregularly distributed derivative estimates. Schaefer and Doswell (1979), Doswell and Caracena (1988), and Spencer and Doswell (2001) explored these two options and concluded that the latter provides superior results.

In diagnostic studies, gridded fields of both observations and spatial derivatives typically are combined in various ways (e.g., advection terms) to evaluate terms in diagnostic and prognostic equations. When gridded spatial derivatives obtained by analyzing irregularly distributed derivative estimates are combined with analyses of observations in order to compute terms in various equations, an inconsistency arises—analyses of the observations are not mutually consistent with their respective gradient analyses. This is because the gradient analyses are not derived from the gridded observations. SSF03 resolved this by developing a simple variational analysis scheme that combines the analysis of a scalar variable with the analysis of its gradient to produce mutually consistent scalar and gradient analyses. SSF03 found that improved analyses are possible when gradient information is included properly within the variational scheme.

In this work, we attempt to improve upon the work of SSF03 by developing a variational scheme that does not require analyses of the observations and its gradient. The scheme that we propose, however, is very similar to that of SSF03, except that our scheme requires only the observations themselves and gradient estimates at points between observations. We compare analyses from our scheme to those from three others (a popular version of the Barnes two-pass, a Barnes three-pass, and SSF03's variational scheme) to evaluate the merits of the proposed variational scheme.

In section 2, we describe our method for generating pseudo-observations that serve as input for the analysis schemes. Section 3 contains descriptions of the various analysis schemes that are to be compared. In section 4, we present the results of our comparisons. Finally, section 5 contains a summary of the results and some brief discussion.

2. Generating pseudo-observations and artificial observing networks

Many previous studies of objective analysis techniques incorporate analytic observations that consist of various combinations of sine and cosine functions, often creating checkerboard patterns in the scalar field when two-dimensional analyses are performed (e.g., Askelson et al. 2000; Trapp and Doswell 2000; Spencer and Doswell 2001). Generating observations by sampling such a field is useful for objective analysis studies because the true field and its spatial derivatives are known at all points in the domain, including the grid points, where the merits of the analysis scheme are evaluated. Simple analytic fields consisting of checkerboard patterns, however, do not replicate the complex atmospheric structure often seen in meteorological observations. For this reason, we have chosen to define our analytic observations in a different manner.

The observations used in this study are 850-mb heights provided by random sampling of grid points from the Rapid Update Cycle Version 2 model (RUC-2; Benjamin et al. 1998). Treating model data as observations introduces atmospheric structure that is more realistic than the simplistic structure of checkerboard patterns. The RUC-2 analysis grid consists of 151 × 113 grid points in the horizontal, covering the contiguous United States and surrounding areas. Only a small percentage of these 17 063 grid points are treated as observations. The following procedure is used to determine which of the 850-mb-height gridpoint data from a given RUC-2 analysis are to be treated as observations for the objective analysis schemes:

1) *Determine the desired percentage of model grid points to use as observations.* If we wish to create an artificial observing network whose observation density (number of observations per unit area) equals that of the U.S. rawinsonde network, for example, then we need to solve the following equation for PERCENT, the percentage of RUC-2 grid points to serve as observations:

$$\frac{72}{8\,151\,824\ \text{km}^2} = \frac{(\text{PERCENT}/100\%)(151 \times 113)}{23\,440\,000\ \text{km}^2}, \tag{1}$$

where 72 is the number of rawinsonde observations within the contiguous United States, 8 151 824 km^2

[1] Strictly speaking, a background field obtained from prior information is not required. An analysis of the observations themselves may serve as a background field.

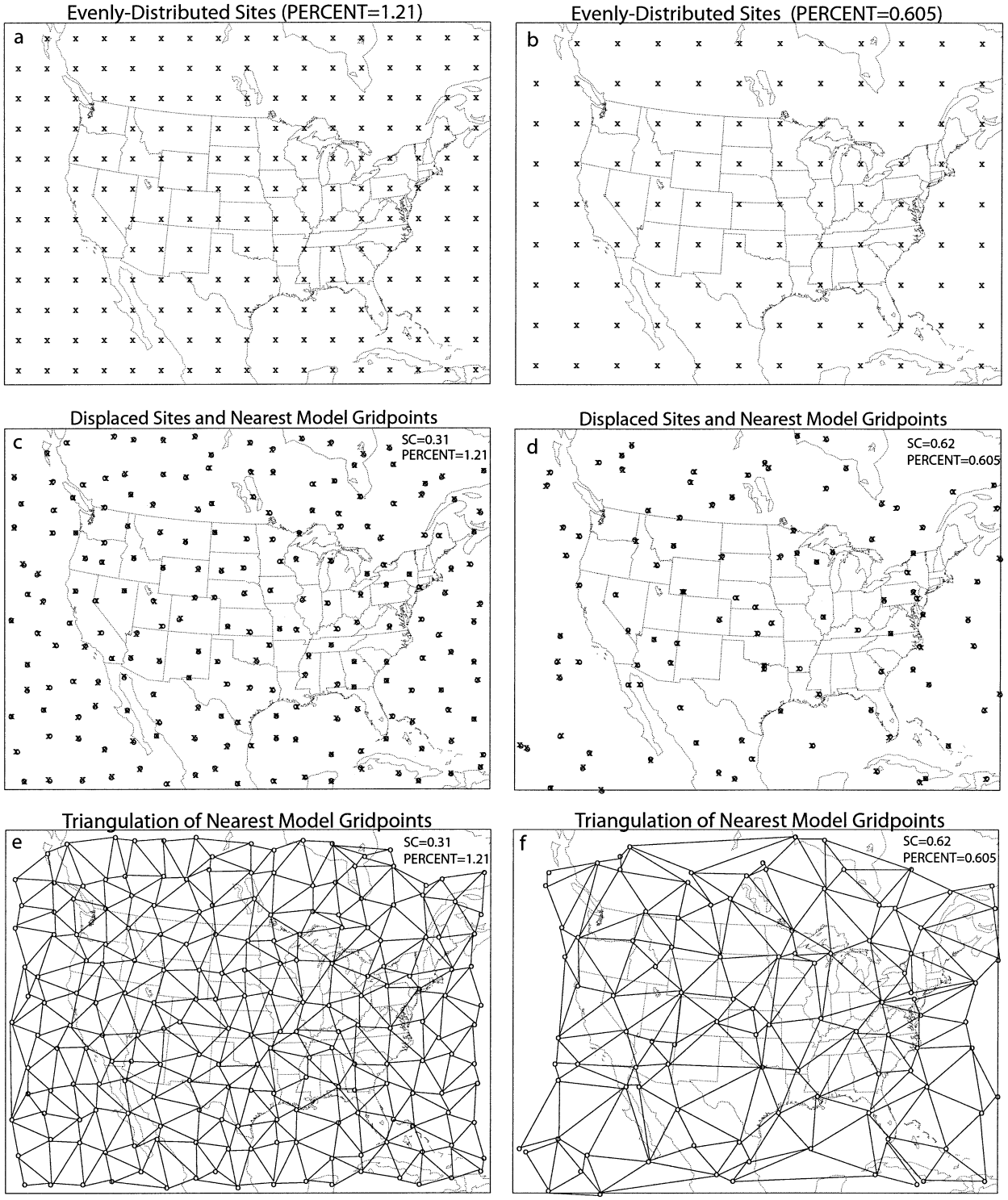

FIG. 1. (a) Artificial network of evenly distributed observing sites created by choosing PERCENT = 1.21; (b) same as (a), except for PERCENT = 0.605; (c) artificial network of irregularly distributed observing sites (denoted by x's) created by choosing SC = 0.31 and PERCENT = 1.21 and nearest model grid points (denoted by o's); (d) same as (c), except for SC = 0.62 and PERCENT = 0.605; (e) triangular tessellation of the nearest model grid points shown in (c); (f) triangular tessellation of the nearest model grid points shown in (d).

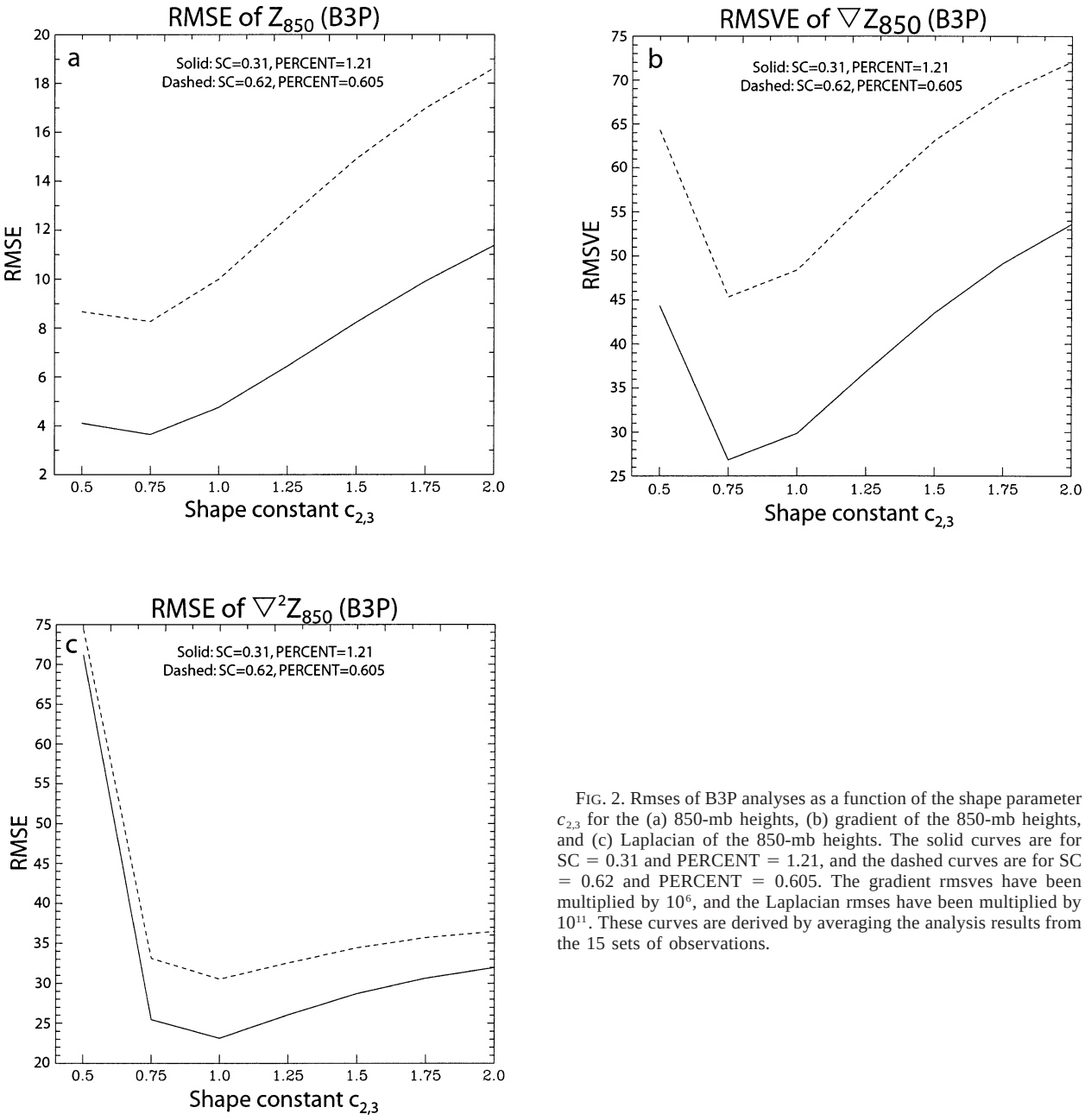

FIG. 2. Rmses of B3P analyses as a function of the shape parameter $c_{2,3}$ for the (a) 850-mb heights, (b) gradient of the 850-mb heights, and (c) Laplacian of the 850-mb heights. The solid curves are for SC = 0.31 and PERCENT = 1.21, and the dashed curves are for SC = 0.62 and PERCENT = 0.605. The gradient rmsves have been multiplied by 10^6, and the Laplacian rmses have been multiplied by 10^{11}. These curves are derived by averaging the analysis results from the 15 sets of observations.

is the approximate area of the contiguous United States, 151 × 113 is the number of RUC-2 grid points in the horizontal, and 23 440 000 km^2 is the approximate area of the RUC-2 domain. Solving (1), we find that PERCENT = 1.21%. Therefore, by choosing 0.0121 × (151 × 113) ≈ 206 RUC-2 grid points to serve as observations, we have mimicked the observation density of the rawinsonde network.

2) *Create a rectangular array of artificial, evenly distributed observing sites across the analysis domain, whose number approximately equals the desired number of observations found in step 1.* To find the exact number of grid points to use as observations and to determine the dimensions of the rectangular array, two equations are solved simultaneously:

$$I \times J = (151 \times 113)(\text{PERCENT}/100\%), \quad (2a)$$

$$I/J = 151/113, \quad (2b)$$

where (I, J) represents the dimension of the rectangular array of artificial, evenly distributed observing sites, 151 × 113 is the horizontal dimension of the RUC-2 domain, and PERCENT is the value found in step 1. Equation (2a) simply states that the

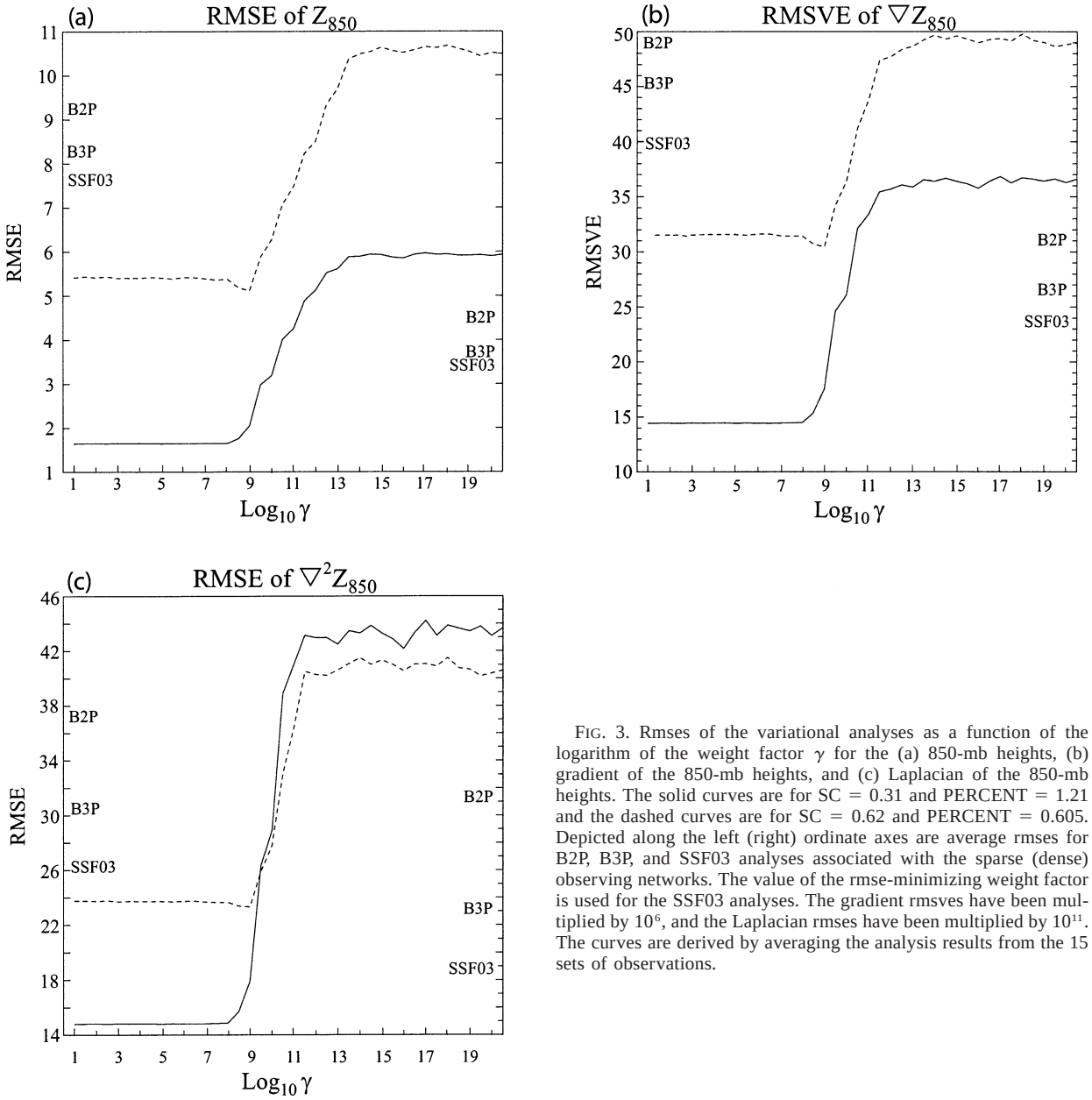

FIG. 3. Rmses of the variational analyses as a function of the logarithm of the weight factor γ for the (a) 850-mb heights, (b) gradient of the 850-mb heights, and (c) Laplacian of the 850-mb heights. The solid curves are for SC = 0.31 and PERCENT = 1.21 and the dashed curves are for SC = 0.62 and PERCENT = 0.605. Depicted along the left (right) ordinate axes are average rmses for B2P, B3P, and SSF03 analyses associated with the sparse (dense) observing networks. The value of the rmse-minimizing weight factor is used for the SSF03 analyses. The gradient rmsves have been multiplied by 10^6, and the Laplacian rmses have been multiplied by 10^{11}. The curves are derived by averaging the analysis results from the 15 sets of observations.

number of artificial observing sites is the desired percentage of the total number of RUC-2 grid points. Equation (2b) states that the aspect ratio of the rectangular array of evenly distributed observing sites should be equivalent to that of the model grid. Solving (2), we find

$$I = 151\left(\frac{\text{PERCENT}}{100}\right)^{0.5}, \tag{3a}$$

$$J = 113\left(\frac{\text{PERCENT}}{100}\right)^{0.5}. \tag{3b}$$

Using the value of PERCENT found in step 1 (i.e., PERCENT = 1.21), we find $(I, J) = (17, 12)$, after rounding to the nearest integer. A 17 × 12 rectangular array of artificial, evenly distributed observing sites is presented in Fig. 1a. This procedure allows us to create an artificial observing network whose observation density mimics that of the rawinsonde network and whose stations are evenly distributed within the analysis domain.

If we wish to create an artificial observing network whose observation density is one-half that of the rawinsonde network, then we choose PERCENT = 0.605. From Eq. (3), we find $(I, J) = (12, 9)$. A 12 × 9 rectangular array of artificial, evenly distributed observing sites is presented in Fig. 1b.

3) *Displace the artificial, evenly distributed observing sites to create an irregularly distributed observing network.* The method described by Doswell and Lasher-Trapp (1997) is used for displacing the evenly distributed observing sites to create a network of irregularly distributed observations. This procedure allows us to mimic the degree of spatial irregularity associated with actual observing networks. To create an observing network that is irregularly distributed, each of the artificial, evenly distributed stations of the rectangular array created in step 2 are randomly displaced in the horizontal by a distance that is no greater than some fraction of Δ, the average data spacing of the evenly distributed sites. This fraction is referred to as the scatter constant (SC). Using SC = 0.0 results in an observing network that remains regularly distributed, whereas using SC = 1.0 generates an observing network that is highly irregularly distributed.[2] If an observing site is displaced far enough that it falls outside any boundary of the domain, then it is merely reflected inside the opposite boundary that same distance in order to keep it within the domain. SSF03 suggest that by using SC = 0.31, an artificial observing network is created whose degree of spatial irregularity matches that of the rawinsonde network. An artificial network created by choosing SC = 0.31 and PERCENT = 1.21 is presented in Fig. 1c. Note that the spatial characteristics (data density and spatial irregularity) of this artificial observing network resemble those of the rawinsonde network, as intended.[3] By doubling the spatial irregularity (SC = 0.62) and halving the observation density (PERCENT = 0.605), the network shown in Fig. 1d is created. This network is characterized by larger data-void regions than is the network shown in Fig. 1c.
4) *For each displaced artificial observing site, choose the nearest model gridpoint datum to serve as an observation.* In this study, observations are created by choosing the 850-mb height at the grid point nearest each displaced observing site. Five different data distributions, each with the same value of SC and PERCENT, are created from each of three RUC-2 analyses (0000 UTC 11 April, 1200 UTC 11 April, and 0000 UTC 12 April 2001)[4] by initializing the random number generator with five different seeds (see footnote 3). Therefore, 15 sets of observations are created for each desired combination of SC and PERCENT. An example of observation locations for SC = 0.31 and PERCENT = 1.21 is presented in Fig. 1c, where the circles represent the nearest model grid points that provide the observations. An example of observation locations for SC = 0.62 and PERCENT = 0.605 is presented in Fig. 1d.

3. Objective analysis schemes

In this section, we describe the three analysis schemes that form the basis of our comparisons. Two of the analysis schemes are of the Barnes type—one is a two-pass and the other is a three-pass successive correction scheme. The third scheme is a modification of the variational scheme described by SSF03. For each of the schemes, the analysis grid is the entire 151 × 113 horizontal RUC-2 grid at the 850-mb level. The nominal grid spacing (Δx) is 40 km.

a. Barnes two-pass scheme

The Barnes (1964, 1973) analysis scheme for objective analysis applies the following analysis equation to the height observations:

$$Z_g = Z_b + \frac{\sum_{k=1}^{N} w_k(Z_k - Z_{\text{bk}})}{\sum_{k=1}^{N} w_k}, \tag{4}$$

where Z_g is the gridded height field, Z_b is the background (first guess) estimate, N is the number of observations, Z_k is the height observation at the kth data point, Z_{bk} is the height estimate at the kth data point obtained by bilinear interpolation of the background field,[5] and w_k is the weight assigned to the kth observation. The weighting function used in (4) is exponential, and we write it as

$$w_k = e^{-R_k^2/(c_i\Delta)^2}, \tag{5}$$

where R_k is the distance between the kth observation and the grid point in question, Δ is the average data spacing of the evenly distributed observations, and c_i is the shape constant used during the ith pass of the analysis. The choices of c_i are made by the analyst and determine the response characteristics of the analysis (i.e., the smoothness of the analysis). Important factors to consider when choosing c_i should include the accuracy and representativeness of the observations,[6] data

[2] Stations within 75 km of each other (a subjectively chosen value) are considered clustered, and the observations associated with these stations are averaged according to the algorithm described by SSF03. Data clustering may occur for moderate to large values of SC.

[3] A completely different distribution of stations, with the same spatial characteristics, may be created simply by altering the value of the initial seed that is used to generate the sequence of random numbers that determine how each station within the evenly distributed, artificial observing network is displaced.

[4] These dates were chosen because a very intense low pressure system developed and moved across the central United States during this period. Also, we note that before any RUC-2 gridpoint data are used as observations, we apply the cowbell filter of Barnes et al. (1996) to the RUC-2 height analyses in order to remove noise.

[5] In our applications of the Barnes scheme (both two- and three-pass), the background field (Z_b) is zero during the first pass. The background field for subsequent passes is simply the analysis from the previous pass.

[6] For this study, the observations are assumed to be free of error.

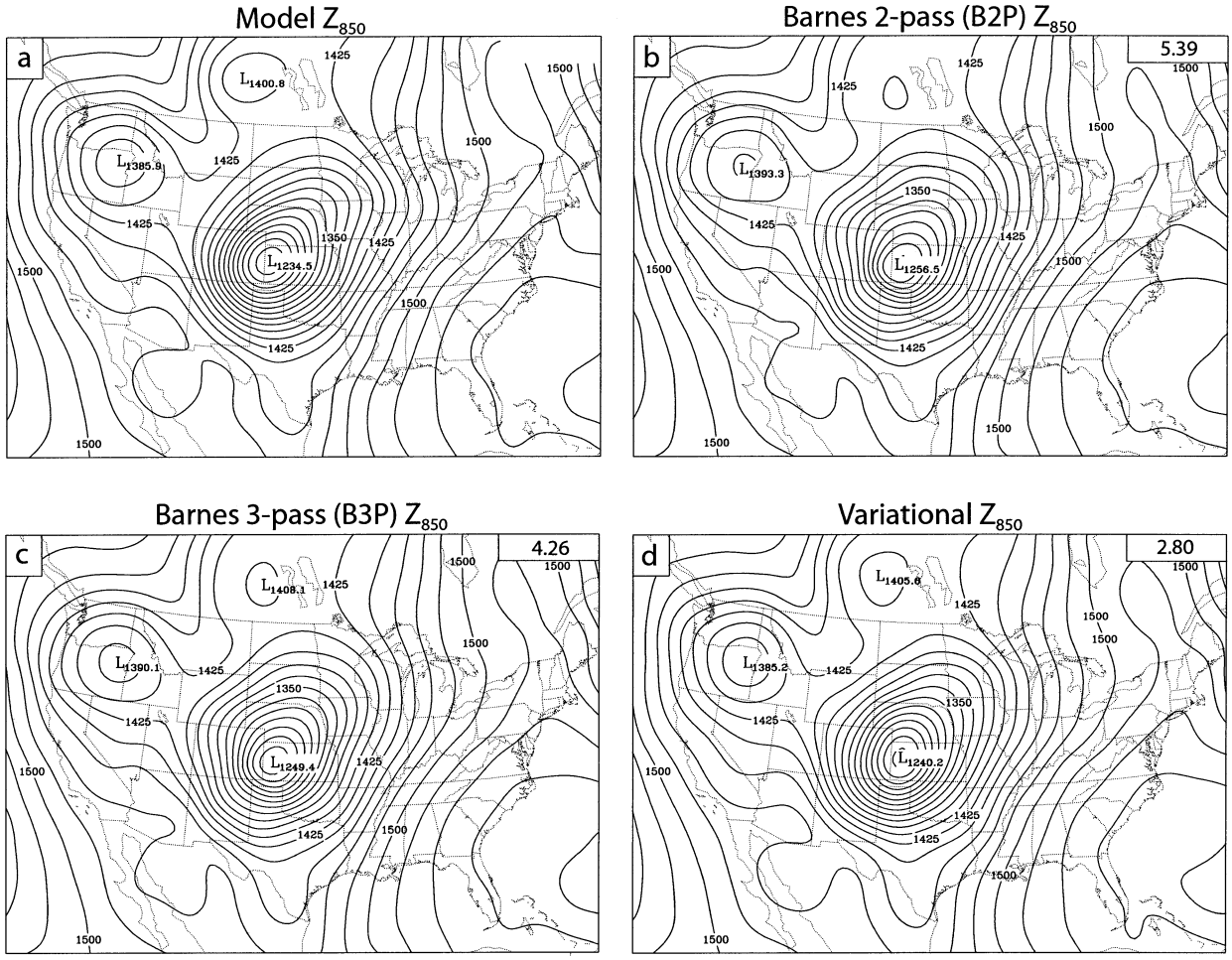

FIG. 4. Sample analyses of the 1200 UTC 11 Apr 2001 850-mb heights from the (a) RUC-2 (''truth''), (b) B2P scheme, (c) B3P scheme, and (d) variational scheme. Pseudo-observations were generated by choosing SC = 0.31 and PERCENT = 1.21 (see Fig. 1c). The contour interval for all plots is 15 m. Rmses (m) appear in the upper-right corners of (b)–(d).

distribution characteristics (Doswell and Lasher-Trapp 1997), and the desired scales to resolve (Barnes 1973).

A two-pass version of the Barnes analysis scheme (hereafter B2P) has been popular for many years for diagnostic studies. In fact, the general meteorological data assimilation, analysis, and display software (GEMPAK) software (Koch et al. 1983), in use today at many universities and research centers, often is used to apply a B2P scheme to meteorological observations. Koch et al. (1983) describe a two-pass scheme in which the first-pass amplitude response at the Nyquist wavelength ($L = 2\Delta$) is 0.0064, and the final response at the Nyquist wavelength is e^{-1}. Using our notation, this may be accomplished by choosing $c_1 = 1.43$ and $c_2 = 0.64$.[7] We have chosen to include this scheme in our comparisons because of the B2P's historical popularity among analysts.

b. Barnes three-pass scheme

The Barnes three-pass scheme (hereafter B3P) is identical to B2P, except that the B3P performs two correction passes through the data rather than the one correction pass that B2P performs. Also, B3P uses values of c_1 that differ from those used by B2P. Achtemeier (1987) showed that improvements in the analysis, particularly for the short but resolvable wavelengths, are possible if a three-pass scheme is used instead of the more popular two-pass scheme. Achtemeier (1989) suggests using a relatively large value of c_1 (first-pass shape constant) so that the first-pass analysis retains significant amplitude only at the well-resolved wavelengths. Following SSF03, we choose $c_1 = 1.75$. This smooth first-pass analysis virtually eliminates the Nyquist wave ($L = 2\Delta$) and retains less than 70% of the amplitude of the largest

[7] These numbers are based on a *one*-dimensional application of the analysis scheme. The shape constants required for accomplishing these response characteristics for a two-dimensional application ($c_1 = 1.01$, $c_2 = 0.45$) are quite low and not recommended.

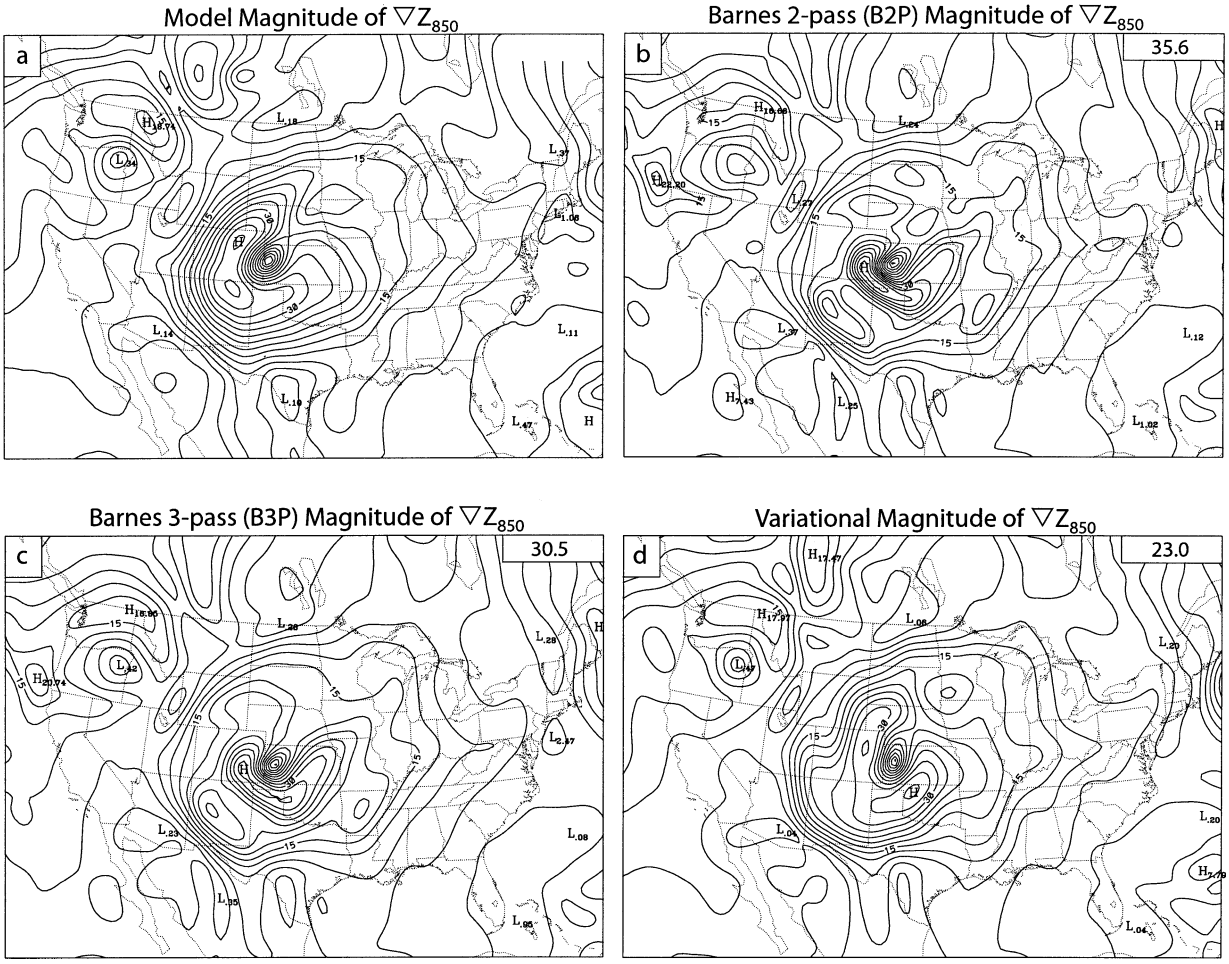

FIG. 5. Same as Fig. 4, except for the magnitude of the gradient of the 850-mb heights. The contour interval for all plots is 3×10^{-5}. The rmsves located in the upper-right corner of (b)–(d) have been multiplied by 10^6.

waves that are considered marginally well sampled ($L = 12\Delta$) as defined by Doswell and Caracena (1988).

The shape constants for the correction passes are chosen such that $c_2 = c_3 \equiv c_{2,3}$. To determine the optimal value for this parameter, different values of $c_{2,3}$, ranging from 0.5 to 2.0, are used by the B3P scheme to analyze the 15 sets of observations created for each combination of SC and PERCENT. The value of $c_{2,3}$ producing the lowest average analysis error [as defined by the root-mean-square error (rmse); section 4a] is used for the comparisons.

Results from the B3P analyses for various values of $c_{2,3}$ are presented in Fig. 2. Two important results are evident: 1) rmses for a network of relatively sparse, moderately irregularly spaced observations are higher than for a dense network of fairly regularly spaced observations, and 2) rmse-minimizing values of $c_{2,3}$ fall within a narrow range of values near 1.0. The rmse-minimizing value of $c_{2,3}$ for the 850-mb heights and its gradient is 0.75 (Figs. 2a and 2b), whereas the rmse-minimizing value of $c_{2,3}$ for the Laplacian is 1.0 (Fig. 2c). For lesser values of $c_{2,3}$, the overfitting of data causes poor derivative information to be forced into the gap between sample points (Barnes 1994; Spencer and Doswell 2001), whereas for larger values of $c_{2,3}$, the analyses tend to be overly damped. For the remainder of this work, we choose $c_{2,3} = 0.75$. For our purposes, the B3P scheme with $c_1 = 1.75$ and $c_{2,3} = 0.75$ is therefore considered the best possible B3P scheme.

c. Variational scheme

The proposed variational scheme is similar in form to that proposed by SSF03. Specifically, we propose minimizing the following two-dimensional cost function J:

$$J = \iint \left\{ [H(Z_v) - Z_o]^2 + \gamma \left[H\left(\frac{\partial Z_v}{\partial x}\right) - \frac{\partial Z^o}{\partial x} \right]^2 + \gamma \left[H\left(\frac{\partial Z_v}{\partial y}\right) - \frac{\partial Z^o}{\partial y} \right]^2 \right\} dx\, dy, \quad (6)$$

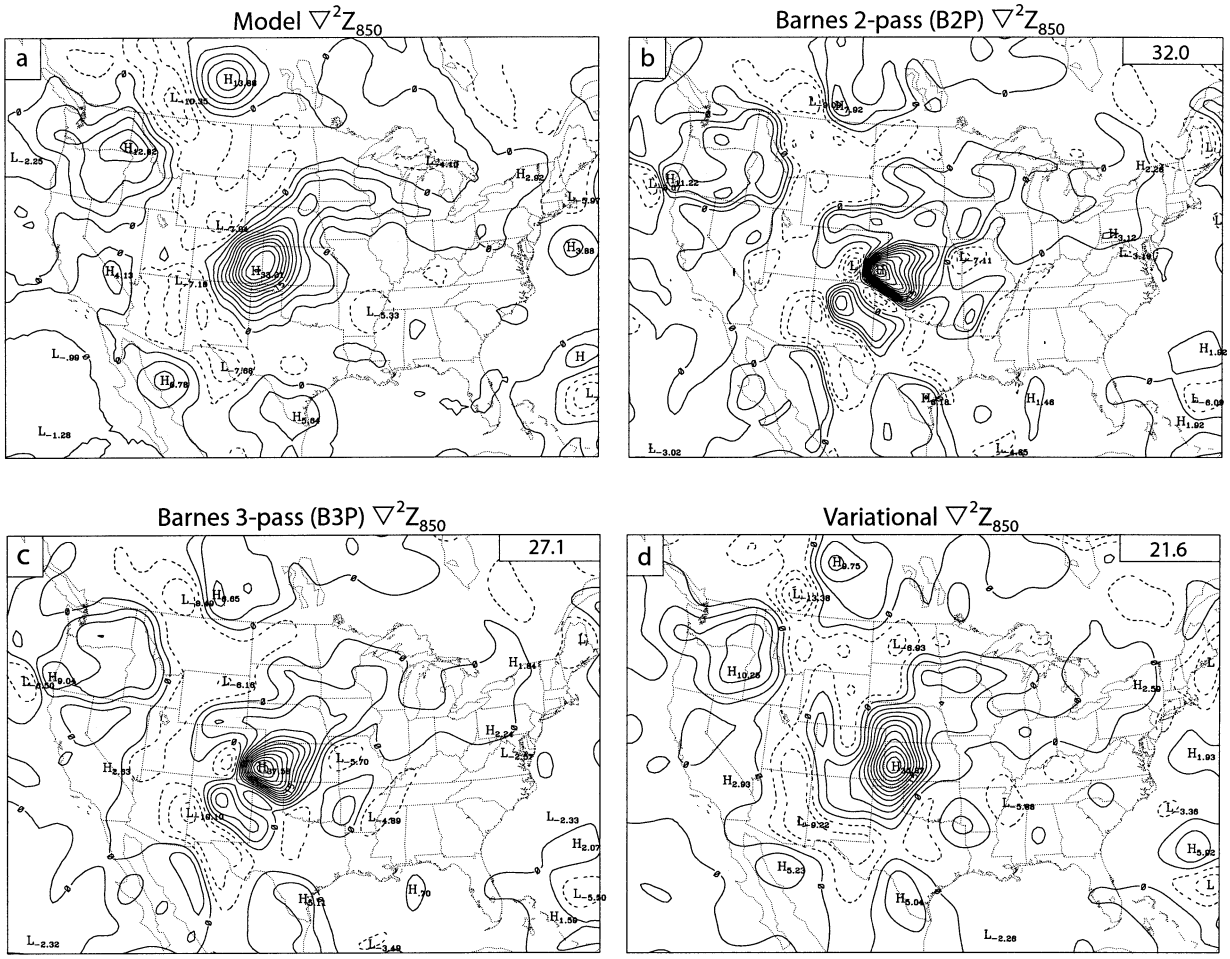

FIG. 6. Same as Fig. 4, except for the Laplacian of the 850-mb heights. The contour interval for all plots is 3×10^{-10} m^{-1}. The rmses located in the upper-right corner of (b)–(d) (in units of m^{-1}) have been multiplied by 10^{11}.

where Z_v represents the desired gridded height field, Z_o are the height observations, $\partial Z^\circ/\partial x$ and $\partial Z^\circ/\partial y$ are the components of the height gradient ''observations,''[8] H is a bilinear interpolation operator that interpolates analysis variables to observation locations, and γ is a user-selectable weight factor that determines the relative strength of the weak constraint. The difference between (6) and the cost function proposed by SSF03 is that their cost function requires explicit interpolation of the scalar observations and its derivatives from irregularly distributed observational points to grid points before the variational analysis is performed, whereas (6) requires actual observations themselves (including gradient observations) and a built-in forward interpolation step (H operator) that interpolates the gridded scalar variable and its gradient to their respective observational locations, where J is computed. As we will demonstrate, this difference in the formulation of the cost function will lead to nontrivial improvements in the analysis of the heights and its derivatives since an analysis routine is not required to determine the terms in (6).

The goal of the variational analysis method is to find the state of Z_v for which the cost function J is minimized. At the minimum, the derivative of J with respect to Z_v vanishes and the optimal estimate of Z_v satisfies

$$\frac{\partial J}{\partial Z_v} = 2\left[\frac{\partial H(Z_v)}{\partial Z_v}\right]^T [H(Z_v) - Z_o] + 2\gamma\left[\frac{\partial H(\partial Z_v/\partial x)}{\partial Z_v}\right]^T\left[H\left(\frac{\partial Z_v}{\partial x}\right) - \frac{\partial Z^o}{\partial x}\right] + 2\gamma\left[\frac{\partial H(\partial Z_v/\partial y)}{\partial Z_v}\right]^T\left[H\left(\frac{\partial Z_v}{\partial y}\right) - \frac{\partial Z^o}{\partial y}\right] = 0. \quad (7)$$

The observations of the gradient are estimated using the ''triangle method'' described in SSF03, which is based on the work of Endlich and Clark (1963). The

[8] These are not true observations, per se. They are *dependent* ''observations'' derived directly from the height observations themselves. Although the gradient ''observations'' are not true observations, they are treated as such.

first step in calculating the gradient of the observations ($\partial Z^\circ/\partial x$, $\partial Z^\circ/\partial y$) is to generate a triangular tesselation of the station locations. We use the Delauney triangulation method described by Ripley (1981) to create a set of nonoverlapping triangles from the observation network. To avoid potential problems with near-colinearity of the vertices, triangles containing minimum angles $\leq 15^\circ$ are removed from further consideration. Examples of triangular tesselations are presented in Figs. 1e and 1f.

SSF03 have shown that if the height varies linearly along each leg of a triangle, then the components of its gradient, calculated for each triangle and assumed valid at the triangle centroid, are given by

$$\frac{\partial Z^o}{\partial x} = \frac{\Delta Z_1}{\Delta x_1} + \left(\frac{\Delta Z_2 \Delta x_1 - \Delta Z_1 \Delta x_2}{\Delta y_1 \Delta x_2 - \Delta y_2 \Delta x_1}\right)\left(\frac{\Delta y_1}{\Delta x_1}\right), \quad (8a)$$

$$\frac{\partial Z^o}{\partial y} = \frac{\Delta Z_2 \Delta x_1 - \Delta Z_1 \Delta x_2}{\Delta y_2 \Delta x_1 - \Delta y_1 \Delta x_2}, \quad (8b)$$

where Z_i is the height observation at the ith station of a triangle ($i \leq 3$), (x_i, y_i) is the location of the ith station, $\Delta Z_i = Z_i - Z_c$, $\Delta x_i = x_i - x_c$, and $\Delta y_i = y_i - y_c$. Here, $Z_c = (Z_1 + Z_2 + Z_3)/3$ is the height estimate at a triangle centroid located at (x_c, y_c). Therefore, in Eq. (6), $\partial Z^\circ/\partial x$ and $\partial Z^\circ/\partial y$ represent gradient ''observations'' at triangle centroids, whereas Z_o represents actual height observations at triangle vertices.

A recursive filter is applied to the analysis in each direction before we calculate the cost function (6). This filter is an effective means for spreading the influence of each observation to nearby grid points, thus ensuring a reasonably smooth analysis. Following the work of Purser and McQuigg (1982), Lorenc (1992), and Hayden and Purser (1995), we define a one-dimensional recursive filter as follows:

$$Y_i = \alpha Y_{i-1} + (1 - \alpha)X_i \quad \text{for } i = 1, \ldots, n,$$

$$Z_i = \alpha Z_{i+1} + (1 - \alpha)Y_i \quad \text{for } i = n, \ldots, 1, \quad (9)$$

where X_i is the initial value at grid point i, Y_i is the value after filtering for $i = 1$ to n, Z_i is the initial value after one pass of the filter in each direction, and α is the filter coefficient given by the following formulation (Lorenc 1992):

$$\alpha = 1 + E - \sqrt{E(E + 2)}, \qquad E = N\Delta x^2/L^2, \quad (10)$$

where L is the horizontal correlation scale, Δx is the grid spacing, and N is the number of filter passes to be applied. In this work, we choose L to be equivalent to the data spacing Δ (~520 km for the sparse networks and ~360 km for the dense networks) and choose $N = 2$. Equation (9) is a first-order recursive filter, applied in both directions to ensure zero phase change. Multipass filters ($N > 1$) are developed by repeated applications of (9). A two-dimensional filter was constructed by applying this one-dimensional filter successively in each coordinate direction. It can be shown that such multidimensional filters, when applied with several passes, can accurately model isotropic Gaussian error correlations (Purser et al. 2003; Gao et al. 2004).

The procedure for solving the variational problem is as follows:

1) Choose a first guess for Z_v, usually zero or a background field obtained from a numerical model.
2) Apply the two-dimensional recursive filter to the guess field and use the forward operator H to interpolate Z_v to triangle vertices and $\partial Z_v/\partial x$, $\partial Z_v/\partial y$ to triangle centroids.
3) Calculate the cost function J using the data at observation locations (triangle vertices and triangle centroids).
4) Calculate the gradient of the cost function at each grid point [Eq. (7)] using the adjoint technique (e.g., Talagrand and Courtier 1987; Courtier et al. 1998).
5) Use a quasi-Newton minimization algorithm (Liu and Nocedal 1989) to obtain updated values of the analysis variable at each grid point as follows:
$$Z_v^{(n)} = Z_v^{(n-1)} + \sigma \cdot f(\partial J/\partial Z_v), \quad (11)$$
where n is the iteration number, σ is the optimal step size obtained by the ''line search'' process in optimal control theory (Gill et al. 1981), and $f(\partial J/\partial Z_v)$ is the optimal descent direction obtained by combining the gradients from several previous iterations.
6) Check whether the optimal solution for Z_v has been found. This is done either by (a) computing the value of J to determine if it is less than a prescribed tolerance or (b) determining if a specified maximum number of iterations has been reached. If either criterion is satisifed, then stop iterating. Otherwise, repeat steps 2 through 6 using the updated field of Z_v as the new guess.

4. Results

a. Measure of analysis error

The rmse between each height analysis (Z_a) and the corresponding model field (Z_m; representing the ''truth'') is calculated according to

$$\text{rmse} = \left[\frac{\sum_{i,j} (Z_a - Z_m)^2}{N}\right]^{0.5}, \quad (12)$$

where N is the number of grid points. The error calculations are performed over the interior 75% of the analysis grid ($N = 9804$) to prevent boundary errors from contaminating the statistic.

Barnes (1994) suggests that the analysis standard is not how closely the analysis replicates the observations, but how accurately the analysis reconstructs the scalar field *and its derivatives.* For this reason, we compute analysis errors for both the height gradient and the Laplacian as well. For the height gradient, the root-mean-square vector error (rmsve) is computed from

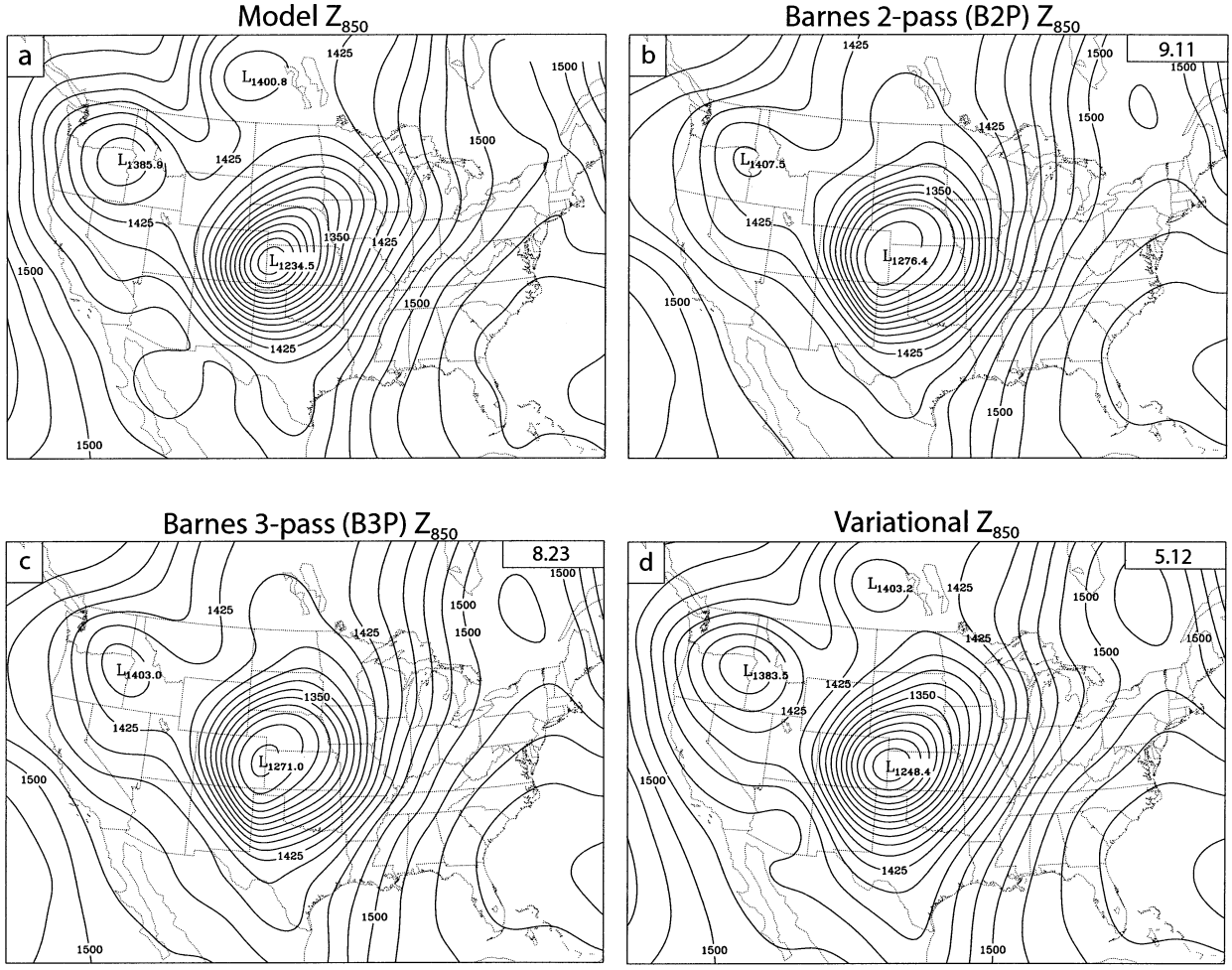

FIG. 7. Same as Fig. 4, except SC = 0.62 and PERCENT = 0.605 (see Fig. 1d).

$$\text{rmsve} = \left\langle \frac{\sum_{i,j} \{[(Z_a)_x - (Z_m)_x]^2 + [(Z_a)_y - (Z_m)_y]^2\}}{N} \right\rangle^{0.5}, \quad (13)$$

where $(Z_a)_x$, $(Z_a)_y$ represent the gridded height gradient components computed from the height analysis Z_a, and $(Z_m)_x$, $(Z_m)_y$ represent the gridded height gradient components computed from the corresponding model field Z_m.[9] For the Laplacian, rmses are given by (12), except that Z_a is replaced by $\nabla^2 Z_a$, and Z_m is replaced by $\nabla^2 Z_m$.

b. *Choosing the weight factor γ for the variational scheme*

The variational formulation [Eq. (6)] contains a weak constraint whose strength is determined by the user-defined weight factor γ. If the analyst wishes that the height observations alone determine the analysis, then a small value of γ is required. On the other hand, if the analyst wants the height gradient observations alone to determine the analysis, then a large value of γ is needed. Using a value of γ somewhere between these two extremes dictates that both the height observations themselves and the gradient information play important roles in the analysis. To choose the value of γ that produces the best analyses for the data under consideration, consider Fig. 3, which plots rmses as a function of the weight factor γ for both sparse, moderately irregularly spaced observations (dashed curves), and for relatively dense, fairly regularly spaced observations (solid curves). Depicted along the left (right) ordinate axis are rmses for B2P, B3P, and SSF03[10] analyses associated with the sparse (dense) observations.

Several points are noted from Fig. 3. First, as with the B3P scheme, we find that the variational scheme produces

[9] Fourth-order centered finite differencing for estimating spatial derivatives is used wherever enough grid points are available. Near the boundaries, second-order centered or one-way finite-differencing schemes are used to estimate derivatives.

[10] The SSF03 variational scheme requires *analyses* of the heights and its gradient. In this work, these are provided by the B3P analyses.

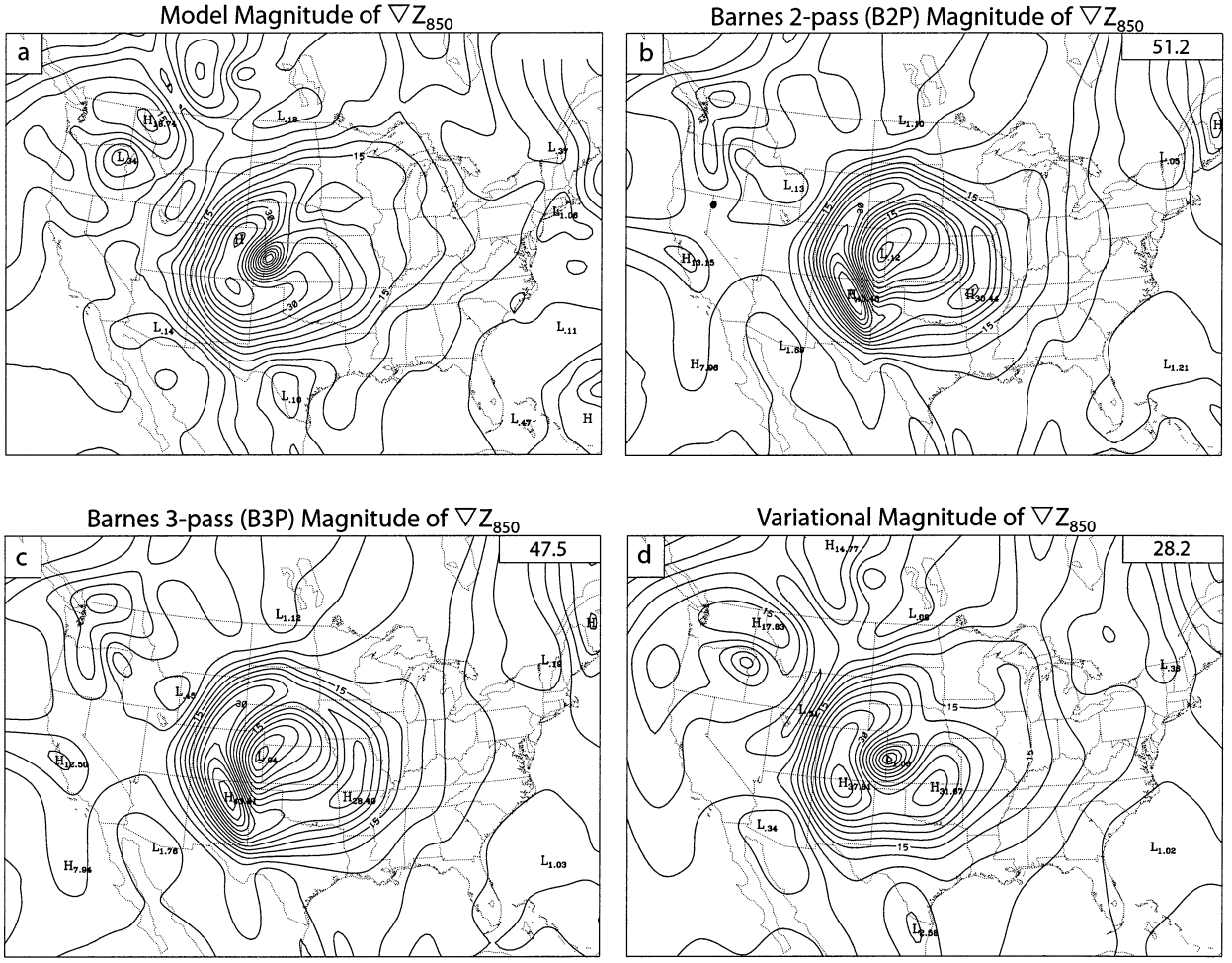

FIG. 8. Same as Fig. 7, except for the magnitude of the gradient of the 850-mb heights. The contour interval for all plots is 3×10^{-5}. The rmsves located in the upper-right corner of each plot have been multiplied by 10^6.

better analyses when the observations are dense and fairly regularly spaced versus when the observations are sparse and somewhat irregularly spaced. This is, of course, no surprise. Second, for both types of data networks under consideration, the use of the height observations alone (small γ) produces much better analyses than when gradient-only observations are used (large γ). Clearly, this too is no surprise. Third, for the dense, fairly regularly spaced observational networks, using height gradient observations in addition to the height observations themselves provides no better analyses than if the height observations alone were used. This is seen by noting that there is no value of γ that provides rmses any lower than the rmses associated with the smallest value of γ on the plots (i.e., $\gamma = 10$ m^2).[11] Therefore, it appears that for these types of observational networks, the gradient observations provide no additional useful information to that which is already available from the height observations themselves. For the sparse, moderately irregularly spaced observations, however, the use of an appropriate value of γ does, in fact, provide additional useful information. For example, when $\gamma = 10^9$ m^2, improvements in the rmses over analyses that use only height observations are 5.6% for the heights (Fig. 3a), 3.8% for the gradient of the heights (Fig. 3b), and 1.9% for the Laplacian of the heights (Fig. 3c). Finally, Fig. 3 indicates that even if the gradient observations are not used, the variational method provides analyses that are superior to those provided by B2P, B3P, and SSF03's variational scheme. This is true for both types of data networks under consideration. For the remainder of this work, γ is held constant at 10^9 m^2.

c. Sample analyses

Sample analyses of the 850-mb heights, magnitude of the height gradient, and Laplacian of the heights from each of three analysis methods for the data distributions shown in Figs. 1e and 1f are presented in Figs. 4–6 and

[11] Using the small value of $\gamma = 10$ m^2 is essentially the same as ignoring the gradient information within the variational scheme.

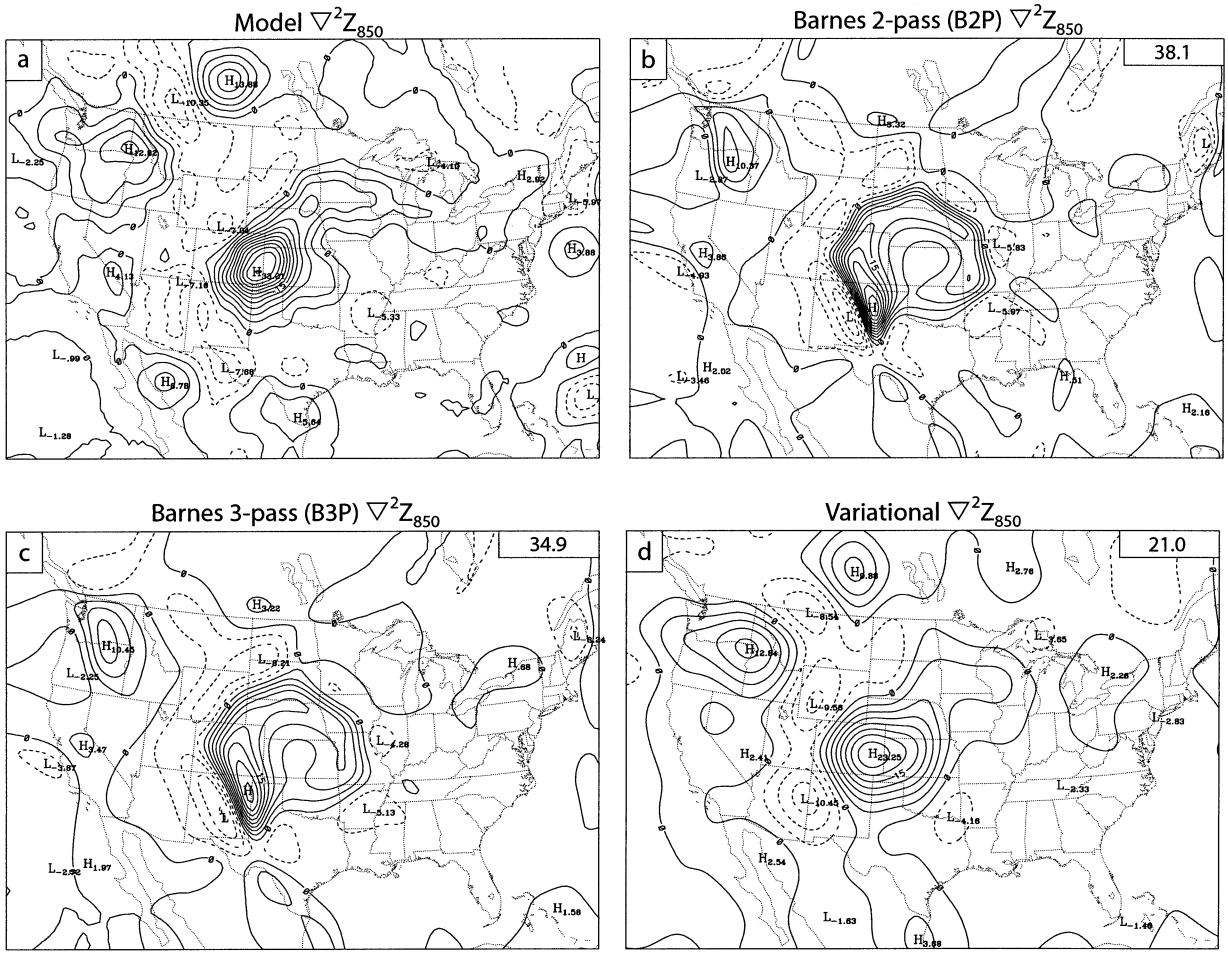

FIG. 9. Same as Fig. 7, except for the Laplacian of the 850-mb heights. The contour interval for all plots is 3×10^{-10} m^{-1}. The rmses located in the upper-right corner of each plot (in units of m^{-1}) have been multiplied by 10^{11}.

7–9, respectively. For all analyses, the rmses (shown in the upper-right corner of each plot) from the B2P scheme exceed those from the B3P scheme which, in turn, exceed those from the variational scheme. Clearly, the variational method is better able to capture the amplitude of the cyclone than are the B2P and B3P schemes (Figs. 4 and 7). For the dense network of observations, the height minimum within the cyclone from the variational method is more than 16 m deeper than that of the B2P scheme and over 9 m deeper than the B3P cyclone (Fig. 4). When the observations are more sparse and irregularly spaced, the differences are 28 and 22.6 m, respectively (Fig. 7).

Visual differences in the patterns of the analyses become even more evident when spatial derivatives are taken (Figs. 5–6 and 8–9). Although differences in the height gradient analyses are obvious (Figs. 5 and 8), differences are especially visually noteworthy for the Laplacian analyses (Figs. 6 and 9). As with the height analyses, differences in the gradient and Laplacian analyses become more pronounced as the observations become fewer and more irregularly spaced.

d. Sensitivity to the spatial irregularity of the observations

We have thus far considered two types of observational networks, one a network of relatively dense observations (PERCENT = 1.21) whose data distribution has been characterized as fairly regular (SC = 0.31)[12] and the other a network of sparse observations (PERCENT = 0.605) whose data distribution has been characterized as somewhat irregular (SC = 0.62). The effects of a variable scatter constant have been isolated by varying SC while holding PERCENT constant. The results for two different values of PERCENT (1.21 and 0.605) are presented in Fig. 10. For all values of SC considered, the B2P scheme provides the poorest analyses and the variational method provides the best analyses. The slopes of the curves suggest that the B2P and B3P schemes are more sensitive to the spatial irregularity of the observations than is the variational method.

[12] Recall that these values of SC and PERCENT allow us to mimic the spatial characteristics of the U.S. rawinsonde network.

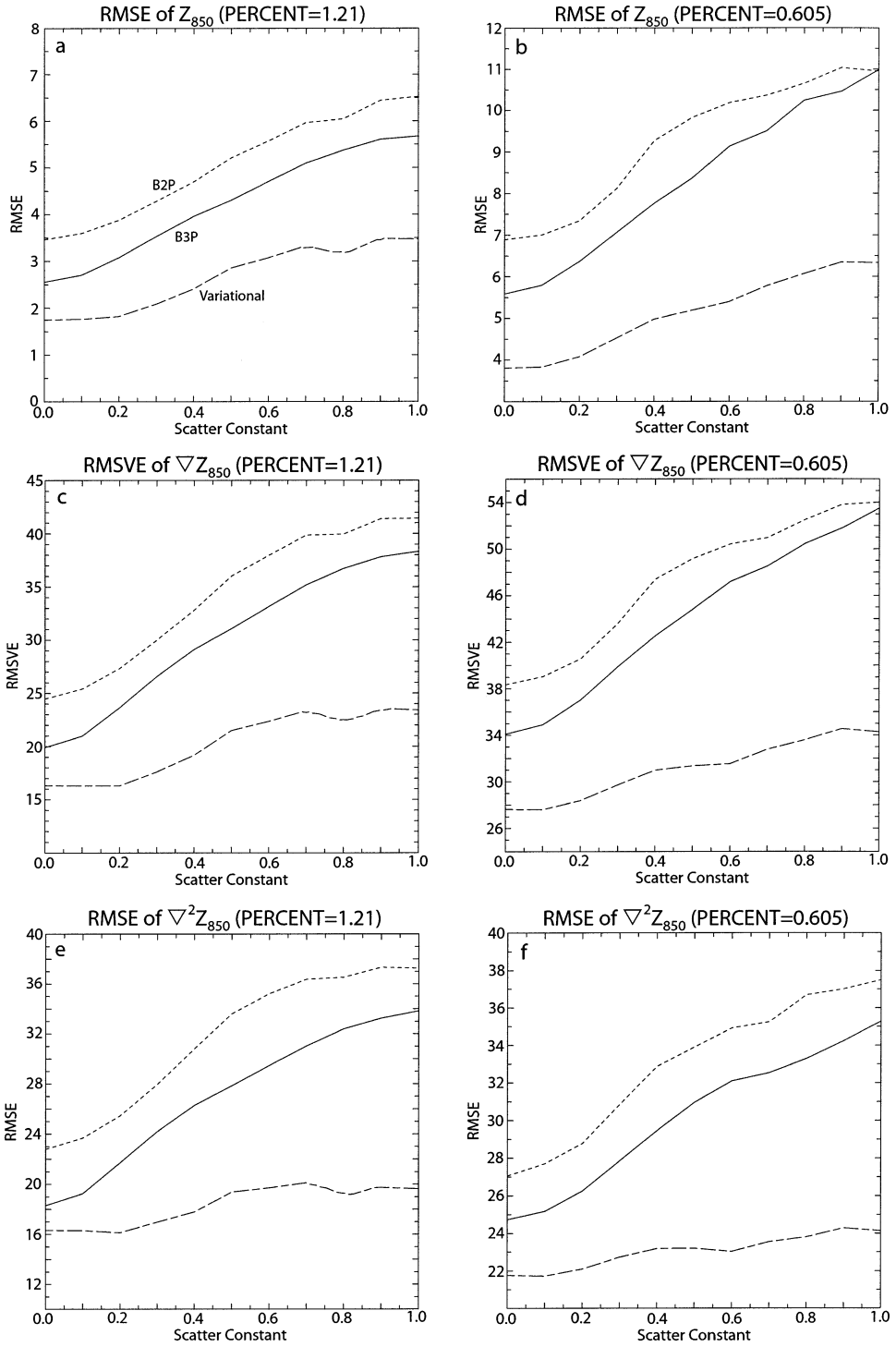

FIG. 10. Rmses of the B2P (short-dashed curves), B3P (solid curves), and variational (long-dashed curves) analyses as a function of the scatter constant. Errors for (a) the 850-mb heights, (c) its gradient, and (e) its Laplacian for PERCENT = 1.21. (b), (d), (f) Errors of the same quantities for PERCENT = 0.605. The gradient rmsves have been multiplied by 10^6, and the Laplacian rmses have been multiplied by 10^{11}. These curves are derived by averaging the analysis results from the 15 sets of observations.

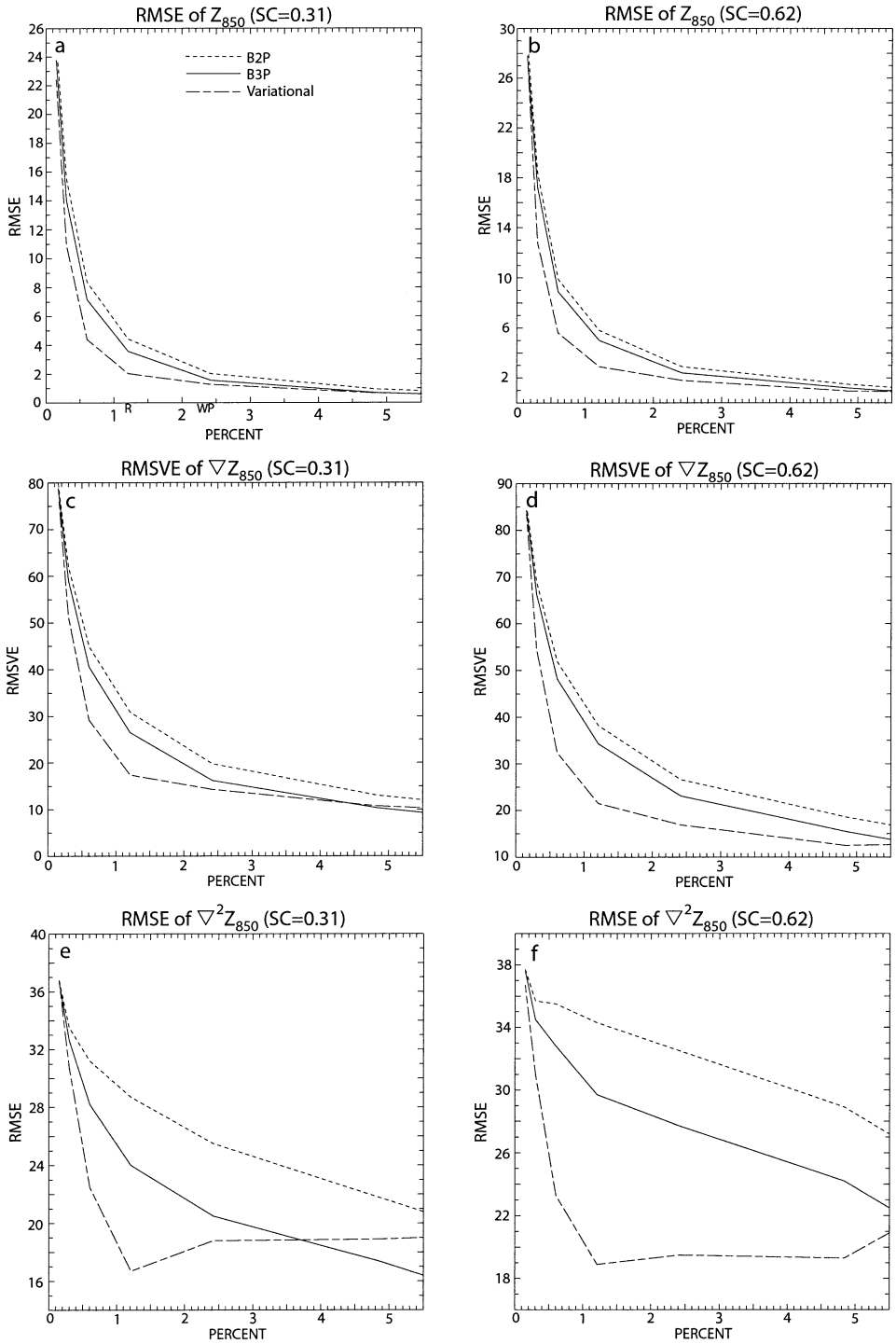

FIG. 11. Rmses of the B2P (short-dashed curves), B3P (solid curves), and variational (long-dashed curves) analyses as a function of PERCENT. Errors for (a) the 850-mb heights, (c) its gradient, and (e) its Laplacian for SC = 0.31. (b), (d), (f) Errors of the same quantities for SC = 0.62. The gradient rmsves have been multiplied by 10^6, and the Laplacian rmses have been multiplied by 10^{11}. These curves are derived by averaging the analysis results from the 15 sets of observations. The letters R, W, and P along the abscissa in (a) represent the values of PERCENT needed to mimic the data densities of the rawinsonde, WSR-88D, and profiler networks, respectively.

In other words, as the observations become more irregularly distributed, the improvements of the variational analyses over the B2P and B3P analyses increase. For example, when PERCENT = 0.605, the rmse of the heights for the variational method is 3.1 m lower than the B2P rmse and 1.8 m lower than the B3P rmse when the data are uniformly distributed (SC = 0.0; Fig. 10b). However, when the observations are highly irregularly spaced (SC = 1.0), the rmse differences are 4.7 m.

e. Sensitivity to the number of observations

We now consider the sensitivity of the analysis methods to the data density by varying the number of observations while holding the scatter constant fixed. The results for two different values of SC (0.31 and 0.62) are presented in Fig. 11. For the values of PERCENT considered, the B2P scheme generally provides the poorest analyses and the variational method provides the best analyses. When very few observations are used (i.e., when PERCENT is very low), the analysis schemes provide essentially equivalently poor analyses. On the other hand, as the number of observations gets very large, the rmses generally tend to converge toward equivalently low values. The largest rmse differences between the analyses lie between these two extremes. These intermediate values of PERCENT characterize many of our current observational networks. For example, we have already described how using a value of PERCENT = 1.21 and SC = 0.31 creates an observational network that mimics the rawinsonde network. Using the results of SSF03 and the procedure described in section 2 for finding PERCENT, we find that the spatial characteristics of the 32-station wind profiler network may be mimicked by selecting PERCENT = 2.43 and SC = 0.25 and that the spatial characteristics of the Weather Surveillance Radar-1988 Doppler (WSR-88D) network may be mimicked by selecting PERCENT = 2.38 and SC = 0.35. These values of PERCENT for the rawinsonde (R), wind profiler (P), and WSR-88D (W) networks are plotted along the abscissa of Fig. 11a. The point of this is to demonstrate that nontrivial differences in the analyses from the three schemes occur for data distributions that resemble those of current observational networks. Therefore, when data from real observational networks are to be analyzed, the choice of an analysis scheme becomes an important issue.

5. Summary and discussion

Comparisons of a new variational technique with traditionally popular Barnes techniques for the analysis of scalar variables has been performed. The variational formulation is very similar to that proposed by SSF03, the difference being that our cost function does not require analyses of the scalar observations and its gradient; the cost function proposed herein simply requires the scalar observations themselves and estimates of the gradient at points between observations (namely, triangle centroids). Another important difference between our work and that of SSF03 is that we have used pseudo-observations that mimic real atmospheric structure rather than appealing to simple, analytically generated checkerboard patterns for the creation of observations. We have done this by randomly picking model gridpoint data to serve as observations. By specifying the percentage of model grid points to serve as observations (through the variable PERCENT) and the degree of irregularity in the spatial distribution of the observations (through the variable SC), we are able to create artificial data networks whose spatial characteristics mimic those of actual observing networks. For example, by choosing PERCENT = 1.21 and SC = 0.31, we were able to create an artificial observing network over the entire RUC-2 domain whose spatial characteristics resemble those of the rawinsonde network. By halving PERCENT and doubling SC, we created an artificial network with one-half the number of observations at twice the spatial irregularity as the rawinsonde network.

Our results suggest that for both types of observational networks considered, the proposed variational scheme produces analyses that are clearly superior to those of the popular Barnes two-pass, ''best'' Barnes three-pass, and SSF03's variational schemes, even when the gradient information is not allowed to influence the analysis (by setting the weight factor γ to a very small value). We find that for a network characterized by relatively dense observations, incorporating the gradient information within the variational scheme did not improve the analysis over that which is available by using only the scalar observations themselves. Evidently, the observations were dense and/or regularly spaced enough so that no benefit was derived by including the gradient information. On the other hand, for a network characterized by relatively sparse observations, including the gradient information does, in fact, improve the analyses if an appropriate value for the weight factor γ is selected. In this work, we found that by choosing $\gamma = 10^9\ \mathrm{m}^2$, we were able to reduce the 850-mb-height rmse by an average of 5.6%, the gradient rmsve by 3.8%, and the Laplacian rmse by 1.9% when compared to analyses created without the use of the gradient information.

We also found that for the data densities considered, the variational scheme produces analyses that are superior to those of the B2P and B3P schemes for all values of the scatter constant that were considered. In other words, the variational scheme performs better whether the data are regularly distributed, highly irregularly distributed, or somewhere in between. In addition, we found that the variational scheme is less sensitive to variations in the irregularity of the spatial distribution of the observations than are the B2P and B3P schemes. Finally, we discovered that the variational scheme outperforms the B2P and B3P schemes for observation

densities that characterize several operational observing networks.

We find it noteworthy that improvements in the variational analyses occur for the sparse data networks and not for the dense data networks when gradient information is allowed to affect the analysis. This is interesting because for a nonlinear distribution of the scalar variable, an increased data spacing implies an increase in the degree of violation of the linearity assumption that is invoked to estimate the scalar gradient. Therefore, as the observational data spacing increases, we expect the gradient estimates to worsen. Evidently, despite the degradation of the gradient estimates as the data spacing increases, the variational scheme still is able to use these estimates to improve upon observation-only analyses, provided that an appropriate choice for the weight factory γ is made. Chien and Smith (1973) have used nonlinear terms in the Taylor series expansion to estimate kinematic quantities from the wind field. They conclude that nonlinear variations may be significant at times and suggest that when the second-order terms of the Taylor series expansion are included, superior results may be achieved. We leave this as an area of future research as it pertains to our variational scheme.

For sparse data networks, we found that by choosing $\gamma = 10^9$ m^2, the average errors of the 850-mb-height analyses from the variational method were minimized. When the 15 sets of analyses are considered individually, the error-minimizing value of γ is almost exclusively in the range 5×10^8 m$^2 \leq \gamma \leq 5 \times 10^9$ m^2.[13] This consistency is important because for larger values of γ, there is a rapid rise in the rmses (Fig. 3). As the value of γ increases beyond about 5×10^9 m^2, the influence of the derivative observations apparently is too strong and the quality of the analyses rapidly worsens.

The proposed variational scheme is intended for diagnostic studies only. As presented, the scheme does not account for observational errors, nor does it include any type of balance constraint. Of course, both of these aspects of data analysis are important when the analysis acts as the initial condition for a numerical model. In the future, we wish to test the concepts developed here to investigate the circumstances (if any) for which model forecasts can be improved by modifying current three-dimensional variational data assimilation (3DVAR) schemes to include the type of gradient observations discussed herein.

Acknowledgments. We appreciate the efforts of Dr. David Stensrud, Dr. Qin Xu, and three anonymous reviewers, whose comments helped us clarify portions of the manuscript. We also wish to thank Mr. Dave Watson and Dr. Chuck Doswell for providing us with the Delauney triangulation code. Major funding for this research was provided under NOAA–OU Cooperative Agreement NA17RJ1227. The second author is supported in part by NSF Grant ATM-0331756.

[13] There are, however, a few individual cases in which the use of gradient observations did not lower the rmses. On average, though, including gradient observations did improve the analyses for the networks characterized by sparse observations.

REFERENCES

Achtemeier, G. L., 1987: On the concept of varying influence radii for a successive corrections objective analysis. *Mon. Wea. Rev.*, **115,** 1760–1771.

——, 1989: Modification of a successive corrections objective analysis for improved derivative calculations. *Mon. Wea. Rev.*, **117,** 78–86.

Askelson, M. A., J. P. Aubagnac, and J. M. Straka, 2000: An adaptation of the Barnes filter applied to the objective analysis of radar data. *Mon. Wea. Rev.*, **128,** 3050–3082.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.*, **3,** 396–409.

——, 1973: Mesoscale objective analysis using weighted time-series observations. NOAA Tech. Memo. ERL NSSL-62, National Severe Storms Laboratory, Norman, OK, 41 pp.

——, 1994: Applications of the Barnes objective analysis scheme. Part I: Effects of undersampling, wave position, and station randomness. *J. Atmos. Oceanic Technol.*, **11,** 1433–1448.

——, F. Caracena, and A. Marroquin, 1996: Extracting synoptic-scale diagnostic information from mesoscale models: The Eta Model, gravity waves, and quasigeostrophic diagnostics. *Bull. Amer. Meteor. Soc.*, **77,** 519–528.

Bellamy, J. C., 1949: Objective calculations of divergence, vertical velocity and vorticity. *Bull. Amer. Meteor. Soc.*, **30,** 45–49.

Benjamin, S. G., J. M. Brown, K. J. Brundage, B. Schwartz, T. Smirnova, T. L. Smith, L. L. Morone, and G. J. DiMego, 1998: The operational RUC-2. Preprints, *16th Conf. on Weather Analysis and Forecasting*, Phoenix, AZ, Amer. Meteor. Soc., 249–252.

Bergthórsson, P., and B. R. Döös, 1955: Numerical weather map analysis. *Tellus*, **7,** 329–340.

Caracena, F., 1987: Analytic approximation of discrete field samples with weighted sums and the gridless computation of field derivatives. *J. Atmos. Sci.*, **44,** 3753–3768.

Ceselski, B. F., and L. L. Sapp, 1975: Objective wind field analysis using line integrals. *Mon. Wea. Rev.*, **103,** 89–100.

Chien, H., and P. J. Smith, 1973: On the estimation of kinematic parameters in the atmosphere from radiosonde wind data. *Mon. Wea. Rev.*, **101,** 252–261.

Courtier, P., and Coauthors, 1998: The ECMWF implementation of three-dimensional variational assimilation (3D-Var). I: Formulation. *Quart. J. Roy. Meteor. Soc.*, **124,** 1783–1807.

Cressman, G., 1959: An operational objective analysis system. *Mon. Wea. Rev.*, **87,** 367–374.

Doswell, C. A., III, and F. Caracena, 1988: Derivative estimation from marginally sampled vector point functions. *J. Atmos. Sci.*, **45,** 242–253.

——, and S. Lasher-Trapp, 1997: On measuring the degree of irregularity in an observing network. *J. Atmos. Oceanic Technol.*, **14,** 120–132.

Endlich, R. M., and J. R. Clark, 1963: Objective computation of some meteorological quantities. *J. Appl. Meteor.*, **2,** 66–81.

Gao, J., M. Xue, K. Brewster, and K. K. Droegemeier, 2004: A three-dimensional variational data analysis method with recursive filter for Doppler radars. *J. Atmos. Oceanic Technol.*, **21,** 457–469.

Gilchrist, B., and G. P. Cressman, 1954: An experiment in objective analysis. *Tellus*, **6,** 309–318.

Gill, P. E., W. Murray, and M. H. Wright, 1981: *Practical Optimization.* Academic Press, 401 pp.

Hayden, C. M., and R. J. Purser, 1995: Recursive filter objective analysis of meteorological fields: Applications to NESDIS operational processing. *J. Appl. Meteor.*, **34,** 3–15.

Koch, S. E., M. desJardins, and P. J. Kocin, 1983: An interactive

Barnes objective map analysis scheme for use with satellite and conventional data. *J. Climate Appl. Meteor.*, **22,** 1487–1503.
Liu, D. C., and J. Nocedal, 1989: On the limited memory BFGS method for large scale optimization. *Math. Programm.*, **45,** 503–528.
Lorenc, A. C., 1992: Iterative analysis using covariance functions and filters. *Quart. J. Roy. Meteor. Soc.*, **118,** 569–591.
——, 1986: Analysis methods for numerical weather prediction. *Quart. J. Roy. Meteor. Soc.*, **112,** 1177–1194.
Panofsky, H. A., 1949: Objective weather-map analysis. *J. Meteor.*, **6,** 386–392.
Parsons, D. B., and J. Dudhia, 1997: Observing system simulation experiments and objective analysis tests in support of the goals of the Atmospheric Radiation Measurement Program. *Mon. Wea. Rev.*, **125,** 2353–2381.
Purser, R. J., and R. McQuigg, 1982: A successive correction analysis scheme using recursive numerical filters. Met. Office Tech. Note 154, British Meteorological Service, 17 pp.
——, W.-S. Wu, D. F. Parrish, and N. M. Roberts, 2003: Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: Spatially homogeneous and isotropic Gaussian covariances. *Mon. Wea. Rev.*, **131,** 1524–1535.
Ripley, B. D., 1981: *Spatial Statistics.* John Wiley and Sons, 252 pp.
Schaefer, J. T., and C. A. Doswell III, 1979: On the interpolation of a vector field. *Mon. Wea. Rev.*, **107,** 458–476.
Spencer, P. L., and C. A. Doswell III, 2001: A quantitative comparison between traditional and line integral methods of derivative estimation. *Mon. Wea. Rev.*, **129,** 2538–2554.
——, D. J. Stensrud, and J. M. Fritsch, 2003: A method for improved analyses of scalars and their derivatives. *Mon. Wea. Rev.*, **131,** 2555–2576.
Talagrand, O., and P. Courtier, 1987: Variational assimilation of meteorological observations with the adjoint vorticity equation. Part I: Theory. *Quart. J. Roy. Meteor. Soc.*, **113,** 1311–1328.
Trapp, R. J., and C. A. Doswell III, 2000: Radar data objective analysis. *J. Atmos. Oceanic Technol.*, **17,** 105–120.
Zamora, R. J., M. A. Shapiro, and C. A. Doswell III, 1987: The diagnosis of upper tropospheric divergence and ageostrophic wind using profiler wind observations. *Mon. Wea. Rev.*, **115,** 871–884.